The Five Factor Model of Personality and Evaluation of Drug Consumption Risk

Elaine Fehrman, Awaz K. Muhammad, Evgeny M. Mirkes, Vincent Egan, and Alexander N. Gorban

Abstract The problem of evaluating an individual's risk of drug consumption and misuse is highly important and novel. An online survey methodology was employed to collect data including personality traits (NEO-FFI-R), impulsivity (BIS-11), sensation seeking (ImpSS), and demographic information. The data set contained information on the consumption of 18 central nervous system psychoactive drugs. Correlation analysis using a relative information gain model demonstrates the existence of a group of drugs (amphetamines, cannabis, cocaine, ecstasy, legal highs, LSD, and magic mushrooms) with strongly correlated consumption. An exhaustive search was performed to select the most effective subset of input features and data mining methods to classify users and non-users for each drug. A number of classification methods were employed (decision tree, random forest, k-nearest neighbours, linear discriminant analysis, Gaussian mixture, probability density function estimation, logistic regression, and naïve Bayes) and the most effective method selected for each drug. The quality of classification was surprisingly high. The best results with sensitivity and specificity being greater than 75% were achieved for cannabis, crack, ecstasy, legal highs, LSD, and volatile substance abuse. Sensitivity and specificity greater than 70% were achieved for amphetamines, amyl nitrite, benzodiazepines, chocolate, caffeine, heroin, ketamine, methadone, and nicotine. The poorest result was obtained for prediction of alcohol consumption.

E. Fehrman (✉)
Men's Personality Disorder and National Women's Directorate, Rampton Hospital, Retford, Nottinghamshire DN22 0PD, UK
e-mail: Elaine.Fehrman@nottshc.nhs.uk

A.K. Muhammad • E.M. Mirkes • A.N. Gorban
Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK
e-mail: akm40@le.ac.uk; em322@le.ac.uk; ag153@le.ac.uk

V. Egan
Department of Psychiatry and Applied Psychology, University of Nottingham, Nottingham NG8 1BB, UK
e-mail: Vincent.Egan@nottingham.ac.uk

F. Palumbo et al. (eds.), *Data Science*, Studies in Classification, Data Analysis, and Knowledge Organization, DOI 10.1007/978-3-319-55723-6_18

1 Introduction

Drug consumption and addiction constitutes a serious problem globally. Drug use is a risk behaviour that does not happen in isolation. It includes numerous risk factors, which are defined as any attribute, characteristic, or event in the life of an individual that increases the probability of drug consumption. Psychological, social, environmental, economic, and individual factors are correlated with initial drug use [8, 43]. These factors are likewise associated with several *personality traits* [5]. Legal drugs such as alcohol, and tobacco are probably responsible for far more premature deaths than illegal recreational drugs [2]. Psychologists have mostly agreed that the personality traits of the Five Factor Model (FFM) are the most comprehensive and adaptable technique for understanding human individual differences [9]. The FFM comprises Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C).

Previous studies demonstrate that high N and O, and low A and C are associated with higher risk of drug use (including cocaine, cannabis, tobacco, heroin, and alcohol) [41]. Our findings improve the knowledge concerning the pathways leading to drug consumption. Sensation seeking (SS) is also higher for users of recreational drugs [24]. The problem of risk evaluation for individuals is much more complex. This was explored very recently [6, 42, 44].

The goal of the study was to predict the risk of drug consumption for each individual according to their personality traits. For this purpose, several data mining approaches were used: decision tree, random forest, k-nearest neighbours, linear discriminant analysis, Gaussian mixture, probability density function estimation, logistic regression, and naïve Bayes. For each drug, the most effective subset of input attributes was selected to provide the highest level of accuracy. Unexpectedly good classifiers were found for all drugs (see Table 1). Successful construction of a classifier provides an instrument for the evaluation of the risk of drug consumption for each individual, along with the creation of a map of risk [12, 31, 32].

The database was collected by an anonymous online survey methodology by Elaine Fehrman yielding 2051 respondents. It included personality traits, impulsivity (Imp), SS, and demographic information including country of location, ethnicity, level of education, gender, and age. The data set contains information on the consumption of 18 central nervous system psychoactive drugs including alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine, and Volatile Substance Abuse (VSA).

Correlation analysis on base of Relative Information Gain (RIG) [33] demonstrates that the consumption of three legal drugs (alcohol, chocolate, and caffeine) is not correlated with other drugs. The consumption of seven illicit drugs (amphetamines, cannabis, cocaine, ecstasy, legal highs, LSD, and magic mushrooms) is symmetrically correlated. There are many strongly asymmetric correlations (see Fig. 3b). For example, knowledge about amphetamines consumption is useful for the evaluation of ketamine usage, but knowledge about ketamine consumption is significantly less useful for the evaluation of amphetamines usage.

2 Materials and Methods

The Database was collected by Elaine Fehrman between 2011 and 2012. An online survey tool from Survey Gizmo was employed to gather data which maximized anonymity, this being particularly relevant to canvassing respondents' views, given the sensitive nature of drug use. All participants were required to declare themselves at least 18 years of age prior to informed consent being given. The study recruited 2051 participants over an 18-month recruitment period. One thousand eight hundred and eighty-five participants (male/female = 943/942) were included following data cleansing. The snowball sampling methodology recruited a primarily (93.7%) native English-speaking sample, with participants from the UK (55.4%), the USA (29.5%), Canada (4.6%), Australia (2.9%), New Zealand (0.3%), and Ireland (1.1%). A total of 6.3% came from a diversity of other countries, none of whom individually met 1% of the sample or did not declare the country of location. Further optimizing anonymity, persons reported their age band, rather than their exact age; 18–24 years (34.1%), 25–34 years (25.5%), 35–44 years (18.9%), 45–54 years (15.6%), 55–64 (4.9%), and over 65 (1%). This indicates that although the biggest age cohort band was in the 18–24 range, some 40% of the cohort was 35 or above, which is an age range often missed in studies of this kind. The sample recruited was highly educated, with 59.5% educated to, at least, degree or professional certificate level: 14.4% reported holding a professional certificate or diploma, 25.5% an undergraduate degree, 15% a master's degree, and 4.7% a doctorate. Approximately 26.8% of the sample had received some college or university tuition although they did not hold any certificates; lastly, 13.6% had left school at the age of 18 or younger. Participants were asked to indicate which racial category was broadly representative of their cultural background. An overwhelming majority (91.2%) reported being White, (1.8%) stated they were Black, and were (1.4%) Asian. The remainder of the sample (5.6%) described themselves as 'Other' or 'Mixed' categories. This small number of persons belonging to specific non-white ethnicities precludes any analyses involving racial categories.

Personality Measurements In order to assess personality traits of the sample, the Revised NEO-Five Factor Inventory (NEO-FFI-R) questionnaire was employed. The NEO-FFI-R was developed and validated by Costa and McCrae [30]. The reliability of its component was studied in [11]. The scale is a 60-item inventory comprised of five personality domains. The five factors are: N, E, O, A, and C with 12 items per domain. Participants were asked to read the 60 NEO-FFI-R statements and indicate on a five-point Likert scale how much a given item applied to them. The second used measure was the *Barratt Impulsiveness Scale* (BIS-11) [40]. The BIS-11 is a 30-item self-report questionnaire, which measures the behavioural

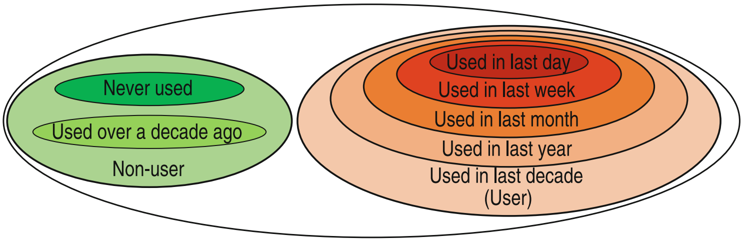

Fig. 1 Categories of drug users

construct of impulsiveness and comprises three subscales: motor impulsiveness, attentional impulsiveness, and non-planning. The scale's items are scored on a four-point Likert scale. The third measurement tool employed was the Impulsiveness Sensation Seeking Scale (ImpSS). Although the ImpSS combines the traits of Imp and SS, it is regarded as a measure of a general SS trait [45]. The scale consists of 19 statements in true–false format, comprising eight items measuring Imp, and 11 items gauging SS.

Drug Use Participants were questioned concerning their use of 18 legal and illegal drugs (alcohol, amphetamines, amyl nitrite, benzodiazepines, cannabis, chocolate, cocaine, caffeine, crack, ecstasy, heroin, ketamine, legal highs, LSD, methadone, magic mushrooms, nicotine, and VSA), and one fictitious drug (Semeron) which was introduced to identify over-claimers (analogously to [21]). Finer distinctions concerning the measurement of drug use have been deployed, due to the potential for the existence of qualitative differences amongst individuals with varying usage levels. It has seven categories of drug users that are depicted in Fig. 1. There are two special categories (see Fig. 1): 'Never used' and 'Used over a decade ago'. These two categories were placed into the class of 'Non-user' and all other five categories into the class 'User', as the simplest version of binary classification.

Input Feature Transformation Many data mining methods were developed to work with continuous data. To use these methods, it is necessary to quantify correctly all the discrete scales and categorical features (for example, 48 categories for each personality trait).

Ordinal Features Quantification The calculation of Polychoric Correlation (PolC) [26] is one of the widely used procedures to analyse categorical data. The matrix of PolC further is used to calculate Principal Components (PCs), etc. The technique of PolC is based on suggestion that categories of ordinal feature are the result of discretization of normally distributed random variable with fixed thresholds. There are two limitations of PolC techniques: it defines the thresholds of discretization but not the values for each category and the defined thresholds are different for different pairs of attributes. Let us have the ordinal feature o with categories o_1,o_2,...,o_k and with number of cases n_i of category o_i. The empirical estimation of probability of category o_i is $p_i = n_i/N$, where $N = \sum n_i$. The sample estimation of thresholds t_i

and average value of i category q_i are evaluating as:

$$t_i = \phi^{-1}\left(\sum_{j=1}^{i} p_j\right), q_i = \phi^{-1}\left(\sum_{j=1}^{i-1} p_j + \frac{p_i}{2}\right).$$

where ϕ is the cumulative normal distribution function.

The correlation coefficients, calculated on base of quantification q_i, have less likelihood than PolC. The merit of this approach is the usage of the same thresholds for all pairs of attributes and explicit formula for calculating the categories' values.

Nominal Feature Quantification We applied the procedure of nonlinear CatPCA [28] to quantify nominal features. This technique includes four steps: exclude nominal features from the set of input features and calculate the informative PCs [14–16, 35] in space of retained input features (to select informative PCs we use Kaiser's rule [18, 23]); calculate the centroid of each category in projection on selected PCs; and calculate the first PC of centroids; put the difference between projection of category centroid and projection of coordinate origin onto calculated PC as the numerical value for this category. As an alternative variant of nominal feature quantification we used dummy coding [17] of nominal variables: country and ethnicity. Each of these attributes was transformed into seven binary features with values 1 or 0.

Input Feature Ranking We used three different procedures of input feature ranking. First, we exploited *principal variables* [29]. Second, we applied Principal Component Analysis (PCA) with *double Kaiser's selection*: Calculated PCs and selected informative PCs by Kaiser's rule [18, 23]. This rule states the all PCs which are corresponding to eigenvalues greater than average are informative and all other PCs are uninformative. We applied the covariance based PCs. For this type of PCs, the Kaiser rule threshold is equal to the sum of diagonal elements of covariance matrix divided by the number of attributes. The importance of an attribute is defined as the maximum of absolute value of the corresponding coordinates in the important PCs represented by unit length vectors. For attribute selection we defined the threshold of importance as $1/\sqrt{n}$ for a unit length vector of dimension n, where n is the number of attributes. If the attribute importance is greater than the threshold of importance, then this attribute is informative. Otherwise the attribute is trivial. If there are trivial attributes, then the worst attribute is the attribute with the minimal value of importance. We removed the worst attribute and repeated the procedure. This procedure stops if there are no trivial attributes. This algorithm ranks attributes from the worst to the best one. Third, we used *sparse PCA* [34]. We applied the simplest threshold for sparse PCA. The detailed description of implemented sparse PCA version is presented in [12]. When using PCA for preselection of the important input features without considering their relationships with drug usage variables, some useful information may be lost. To overcome this limitation of the classical PCA, the goal-oriented tools should be used, like *supervised PCA* [25].

Classification Methods For this study, we applied several classification methods which provided risk evaluation as well.

1. K-Nearest Neighbours (KNN). The basic concept of KNN is: the class of an object is the class of the majority of its k-nearest neighbours [7]. We tested the KNN versions, which differed by: the number of NN, which was varied between 1 and 20; the set of input features; distances such as Euclidean distance, adaptive distance [19], and Fishers distance [13]; the kernel function for adaptive distance transformation; and the kernel functions [27] for voting.
2. Decision Tree (DT) [36, 39] is a method that constructs a tree like structure, which can be used to choose between several courses of action. The binary decision trees were used. We tested the decision trees, which differed by: the three split criterion (information gain, Gini gain, or DKM gain); linearly combined or separately used input features; the set of the input features; and the minimal number of cases in the leaf, which varied between 3 and 30.
3. Linear Discriminant Analysis (LDA). We used Fisher's linear discriminant [13]. We tested the LDA which differed by RIG, Gini gain, DKM gain, or accuracy as criterion for threshold defining.
4. Gaussian mixture is a method to estimate the probability under assumption that each category of target feature has the multivariate normal distribution [10].
5. Probability Density Function Estimation (PDFE). We implemented the radial-basis functions method for it [38]. We tested the PDFE versions which differed by: the number of the nearest neighbours which was varied between 5 and 30; the set of the input features; and the kernel function [27] which was placed in each data points.
6. Logistic regression. We implemented the weighted version of logistic regression [22]. The log likelihood estimation of the regression coefficients was used.
7. Naïve Bayes. We implemented the standard version of naïve Bayes [37]. All attributes which contained less than or equal to 20 different values were interpreted as categorical and the standard contingency tables were calculated for such attributes. All other attributes were considered as normally distributed.
8. Random forest was used for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data [4].

Maximal value of minimum of sensitivity and specificity were employed as the criteria. If minimum of sensitivity and specificity was the same, then we selected classifier with maximal sum of the sensitivity and specificity. This criterion selects the classifier furthest from the 'completely random guess' classifier. Classifiers with sensitivity or specificity less than 50% were not considered. *Leave-One-Out Cross Validation* (LOOCV) [1] was used for all tests.

3 Results

The study's sample was found to be significantly biased in comparison with the general population [11, 30]. Such a bias is usual for clinical cohorts [41]. For each drug average values of personality traits for groups of users are shifted from the sample means in the same direction as the sample means deviate from the population norm. For all drugs the average values of N and O for groups of users are neutral or moderately high. Groups of user of illicit drugs have moderately low A and C. Groups of licit drug users have neutral scores of A and C, apart from nicotine users, who have moderately low scores of C. For the groups of users of crack, heroin, VSA, and methadone, the score of E is moderately low. For groups of users of other drugs, the score of E is neutral. Detailed exploratory analysis of personality traits is presented in [12].

All three feature ranking approaches in Sect. 2 show that ethnicity and country are not important for evaluation of drug consumption risk. The reason for this is the high imbalance of frequency of different categories: more than 91% of cases in the 'White' category of ethnicity and 55% of UK and 29% of USA for country. These two features were removed from further study.

The Pearson's correlation coefficients (PCC) r is not appropriate for general categorical attributes but for the Boolean random variables (with 0,1 values) it gives a reasonable measure of dependence because for them $\mathrm{cov}(X, Y) = P(X = 1 \& Y = 1) - P(X = 1)P(Y = 1)$. The majority of the PCCs between the usage of different drugs are significant. One hundred twenty-four pairs of drug usages from a total of 153 pairs have p-values less than 0.01 (p-value is the probability to observe by chance the same or greater correlation coefficient for independent variables). A multi-testing approach is necessary when testing 153 pairs of drug usages in order to estimate the significance of the correlation [3]. We applied the most conservative technique, Bonferroni correction, and used Benjamini-Hochberg (BH) step-up procedure [3] to control False Discovery Rate (FDR) for estimation of the genuine significance of these correlations. One hundred and fifteen correlation coefficients were significant with Bonferroni corrected p-value 0.001. BH step-up procedure with threshold of FDR equals 0.01 defined 127 significant correlation coefficients. We consider correlations with absolute values of PCC $|r| \geq 0.4$. Figure 2 sets out all identified significant correlations greater than 0.4. The correlation can be considered weak if $|r| < 0.4$; medium if $0.45 > |r| \geq 0.4$; strong if $0.5 > |r| \geq 0.45$; and very strong if $|r| \geq 0.5$. Figure 2 shows that the usage of amphetamines, cannabis, cocaine, ecstasy, ketamine, legal highs, LSD, and magic mushrooms correlates with

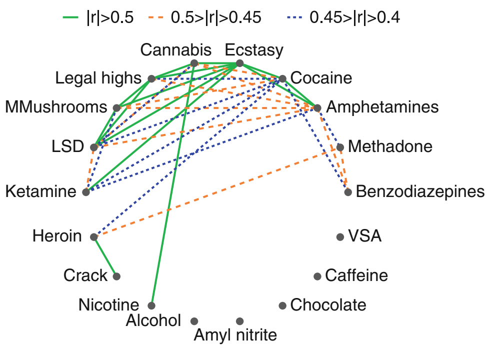

Fig. 2 Strong drug usage correlations

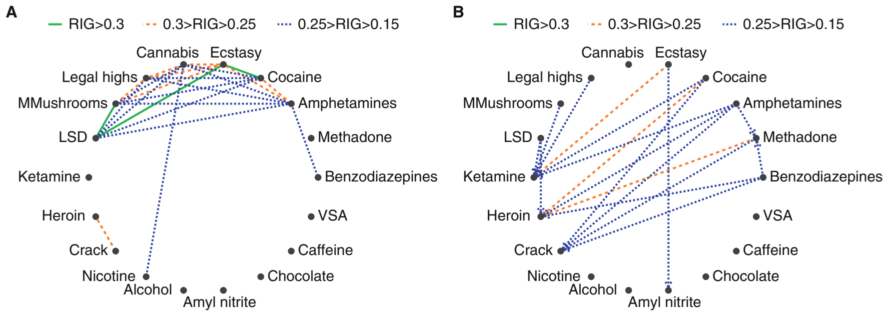

Fig. 3 Pairs of drug usages with high relative information gain: (**a**) more or less symmetric RIG and (**b**) significantly asymmetric RIG. In figure (**b**) *arrow* from LSD usage to heroin usage, for example, means that knowledge of LSD usage can decrease uncertainty in heroin usage

all other drugs in the same group, excluding correlations between cannabis and ketamine usage (r=0.302) and between legal highs and ketamine usage (r=0.393).

RIG is widely used in data mining to measure dependence between categorical attributes [33]. The greater the value of RIG, the stronger is the indicated correlation. RIG is zero for independent attributes. RIG is not symmetric. It is a measure of mutual information. Figure 3 presents all pairs with RIG >0.15. Figure 3a shows "symmetric" RIGs (we call RIG($X|Y$) symmetric if $\frac{|\mathrm{RIG}(X|Y)-\mathrm{RIG}(Y|X)|}{\min\left(\mathrm{RIG}(X|Y),\mathrm{RIG}(Y|X)\right)} < 0.2$). Figure 3b shows asymmetric RIG. It can be seen that in Fig. 3a the usage of amphetamines, cannabis, cocaine, ecstasy, legal highs, LSD, and magic mushrooms are correlated with each other. This group is the same that in Fig. 2 except for ketamine usage. Asymmetric RIGs illustrate pattern significantly different from Fig. 2.

To find the best classifier we used eight different types of classifiers and selected the best one for each drug. Results of the best classifier selection are presented in Table 1. Table 1 shows that for all drugs except alcohol, cocaine, and magic mushrooms, the sensitivity and specificity are greater than 70%. It is an unexpectedly high accuracy. After initial feature selection data contain 10 input features. Each of them is important at least for five drugs. There is no single most effective classifier employing all input features. The maximum number of used features is 6 out of 10 and the least number is 2. Table 1 shows that the best choice of the input attributes is different for different drugs. Age is used in the best classifiers for 14 drugs (the most universal attribute). Gender is used in the best classifiers for 10 drugs.

It is important to note that attributes which are unused in the best classifiers are not non-informative or redundant. For example, for ecstasy the best classifier is based on Age, SS, and Gender and has sensitivity 76.17% and specificity 77.16%. There exist a DT for the same drug based on Age, Edu., O, C, and SS with sensitivity 77.23% and specificity 75.22%, a DT based on Age, Edu., E, O, and A with sensitivity 73.24% and specificity 78.22%, and a KNN classifier based on Age, Edu.,

Table 1 The best results of the drug users classifiers

Target feature	Method	Age	Edu.	N	E	O	A	C	Imp.	SS	Gender	Sens. (%)	Spec. (%)	Sum (%)
Alcohol	LDA	X	X	X						X	X	75.34	63.24	138.58
Amphetamines	DT	X		X		X		X	X	X		81.30	71.48	152.77
Amyl nitrite	DT			X		X		X		X		73.51	87.86	161.37
Benzodiazepines	DT	X		X	X				X	X	X	70.87	71.51	142.38
Cannabis	DT	X	X			X	X	X	X			79.29	80.00	159.29
Chocolate	KNN	X			X			X			X	72.43	71.43	143.86
Cocaine	DT	X				X	X		X	X		68.27	83.06	151.32
Caffeine	KNN	X	X			X	X		X			70.51	72.97	143.48
Crack	DT				X			X				80.63	78.57	159.20
Ecstasy	DT	X								X	X	76.17	77.16	153.33
Heroin	DT	X							X		X	82.55	72.98	155.53
Ketamine	DT	X			X		X		X	X		72.29	80.98	153.26
Legal highs	DT	X				X	X	X		X	X	79.53	82.37	161.90
LSD	DT	X		X	X	X			X		X	85.46	77.56	163.02
Methadone	DT	X	X		X	X					X	79.14	72.48	151.62
MMushrooms	DT				X						X	65.56	94.79	160.36
Nicotine	DT			X	X			X			X	71.28	79.07	150.35
VSA	DT	X	X		X		X	X		X		83.48	77.64	161.12

Symbol 'X' means used input feature. Results are calculated by LOOCV

N, E, O, C, Imp., SS, and Gender with sensitivity 75.63% and specificity 75.75%. It means that for ecstasy users risk evaluation all input attributes are informative but required information can be extracted from part of attributes.

The results presented in Table 1 were calculated by LOOCV. It should be stressed that different methods of testing give different sensitivity and specificity. For example, a decision tree formed for the entire sample can have accuracy, sensitivity, and specificity different from LOOCV [20]. For illustration we can use the decision tree for ecstasy, depicted in Fig. 4. It has sensitivity 78.56% and specificity 71.16%, calculated using the whole sample. Results of LOOCV for a tree with the same options presented in the Table 1 show sensitivity 76.17% and specificity 77.16%.

4 Discussion

We evaluated the individual drug consumption risk for each drug. We analysed interrelations between the individual drug consumption risks for different drugs. We applied eight data mining approaches and selected the best one for each drug. Classifiers with sensitivity and specificity greater than 70% were found for all drugs

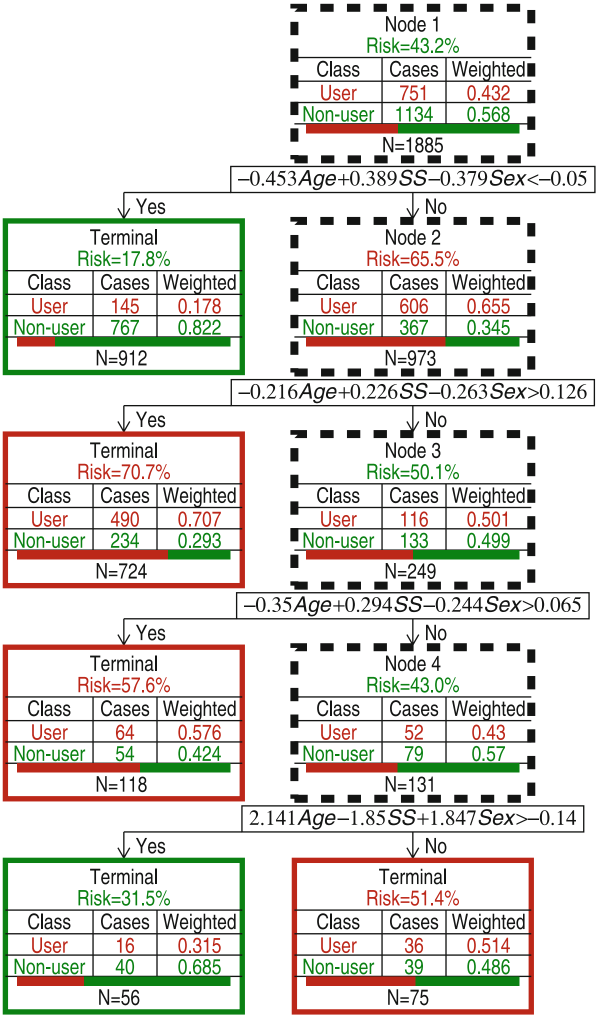

Fig. 4 Decision tree for ecstasy. Input features are: Age, SS, and Gender. Non-terminal nodes are depicted with *dashed border*. Values of Age, SS, and Gender are calculated by quantification procedures described in Sect. 2. Weight of each case of user class is 1.15 and of non-user class is 1. Column 'Weighted' presents normalized weights: weight of each class is divided by sum of weights

except magic mushrooms, alcohol, and cocaine. This accuracy is unexpectedly high for this type of problem. Correlation analysis using a RIG model demonstrated the existence of a group of drugs (see Fig. 3a) with strongly correlated consumption. There are limitations of this study. The collected sample is biased with respect to the general population, but it can still be used for risk evaluation. A further limitation concerns the fact that a number of the findings may be culturally specific.

References

1. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)
2. Beaglehole, R., Bonita, R., Horton, R., Adams, C., Alleyne, G., Asaria, P., et al.: Priority actions for the non-communicable disease crisis. Lancet **377**(9775), 1438–1447 (2011)
3. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Royal Stat. Soc. **57**(1), 289–300 (1995)
4. Biau, G.: Analysis of a random forests model. J. Mach. Learn. Res. **13**(1), 1063–1095 (2012)
5. Bogg, T., Roberts, B.W.:. Conscientiousness and health-related behaviors: a meta-analysis of the leading behavioral contributors to mortality. Psychol. Bull. **130**(6), 887–919 (2004)
6. Bulut, F., Bucak, İ.Ö.: An urgent precaution system to detect students at risk of substance abuse through classification algorithms. Turk. J. Electr. Eng. Comput. Sci. **22**(3), 690–707 (2014)
7. Clarkson, K.L.: Nearest-neighbor searching and metric space dimensions. In: Shakhnarovich, G., Darrell, T., Indyk, P. (eds.) Nearest-Neighbor Methods for Learning and Vision: Theory and Practice, pp. 15–59. MIT, Cambridge (2005)
8. Cleveland, M.J., Feinberg, M.E., Bontempo, D.E., Greenberg, M.T.: The role of risk and protective factors in substance use across adolescence. J. Adolesc. Health **43**(2), 157–164 (2008)
9. Costa, P.T., MacCrae, R.R.: Revised NEO-Personality Inventory (NEO-PI-R) and the NEO-Five Factor Inventory (NEO-FFI): Personality manual. Psychological Assessment Resources, Odessa, FL (1992)
10. Dinov, I.D.: Expectation maximization and mixture modeling tutorial. UCLA, Statistics Online Computational Resource (2008). http://escholarship.org/uc/item/1rb7097
11. Egan, V., Deary, I., Austin, E.: The NEO-FFI: emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. Personal. Individ. Differ. **29**(5), 907–920 (2000)
12. Fehrman, E., Muhammad, A.K., Mirkes, E.M., Egan, V., Gorban, A.N.: The five factor model of personality and evaluation of drug consumption risk. arXiv preprint arXiv:1506.06297
13. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**(2), 179–188 (1936)
14. Gorban, A.N., Zinovyev, A.Y.: Principal graphs and manifolds. In Olivas, E.S., Guerrero, J.D.M., Sober, M.M., Benedito, J.R.M., López, A.J.S. (eds.) Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques, pp 28–59. IGI Global, Hershey, NY (2009)
15. Gorban, A.N., Zinovyev, A.Y.: Principal manifolds and graphs in practice: from molecular biology to dynamical systems. Int. J. Neural Syst. **20**(3), 219–232 (2010)
16. Gorban, A.N., Kégl, B., Wunsch, D.C., Zinovyev, A.Y. (eds.): Principal Manifolds for Data Visualisation and Dimension Reduction. Lecture Notes in Computer Science and Engineering, vol. 58. Springer, Berlin, Heidelberg (2008)
17. Gujarati, D.N.: Basic Econometrics, 4th edn. McGraw-Hill, New York (2003)
18. Guttman, L.: Some necessary conditions for common-factor analysis. Psychometrika **19**(2), 149–161 (1954)
19. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification. IEEE Trans. Pattern Anal. Mach. Intell. **18**(6), 607–616 (1996)
20. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2009)
21. Hoare, J., Moon, D.: Drug misuse declared: findings from the 2009/10 British Crime Survey Home Office Statistical Bulletin 13/10 (2010)
22. Hosmer, D.W. Jr., Lemeshow, S.: Applied Logistic Regression. Wiley, New York (2004)
23. Kaiser, H.F.: The application of electronic computers to factor analysis. Educ. Psychol. Meas. **20**, 141–151 (1960)

24. Kopstein, A.N., Crum, R.M., Celentano, D.D., Martin, S.S.: Sensation seeking needs among 8th and 11th graders: characteristics associated with cigarette and marijuana use. Drug Alcohol Depend. **62**(3), 195–203 (2001)
25. Koren, Y., Carmel, L.: Robust linear dimensionality reduction. IEEE Trans. Vis. Comput. Graph. **10**(4), 459–470 (2004)
26. Lee, S.Y., Poon, W.Y., Bentler, P.M.: A two-stage estimation of structural equation models with continuous and polytomous variables. Br. J. Math. Stat. Psychol. **48**(2), 339–358 (1995)
27. Li, Q., Racine, J.S.: Nonparametric Econometrics: Theory and Practice. Princeton University Press, Princeton, NJ (2007)
28. Linting, M., van der Kooij, A.: Nonlinear principal components analysis with CATPCA: a tutorial. J. Pers. Assess. **94**(1), 12–25 (2012)
29. McCabe, G.P.: Principal variables. Technometrics **26**(2), 137–144 (1984)
30. McCrae, R.R., Costa, P.T.: A contemplated revision of the NEO Five-Factor Inventory. Personal. Individ. Differ. **36**(3), 587–596 (2004)
31. Mirkes, E.M., Alexandrakis, I., Slater, K., Tuli, R., Gorban, A.N.: Computational diagnosis and risk evaluation for canine lymphoma. Comput. Biol. Med. **53**, 279–290 (2014)
32. Mirkes, E.M., Alexandrakis, I., Slater, K., Tuli, R., Gorban, A.N.: Computational diagnosis of canine lymphoma. J. Phys. Conf. Ser. **490**(1), 012135 (2014). http://stacks.iop.org/1742--6596/490/i=1/a=012135
33. Mitchell, T.M.: Machine learning. 1997. Burr Ridge, IL: McGraw Hill **45** (1997).
34. Naikal, N., Yang, A.Y., Sastry, S.S.: Informative feature selection for object recognition via sparse PCA. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 818–825. IEEE, New York (2011)
35. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. **2**(6), 559–572 (1901)
36. Quinlan, J.R.: Simplifying decision trees. Int. J. Man Mach. Stud. **27**(3), 221–234 (1987)
37. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, Prentice Hall, NJ (1995)
38. Scott, D.W.: Multivariate Density Estimation: Theory, Practice and Visualization. Wiley, New York (1992)
39. Sofeikov, K.I., Tyukin, I.Y., Gorban, A.N., Mirkes, E.M., Prokhorov, D.V., Romanenko, I.V.: Learning optimization for decision tree classification of non-categorical data with information gain impurity criterion. In: 2014 International Joint Conference on Neural Networks (IJCNN), pp. 3548–3555. IEEE, New York (2014)
40. Stanford, M.S., Mathias, C.W., Dougherty, D.M., Lake, S.L., Anderson, N.E., Patton, J.H.: Fifty years of the Barratt impulsiveness scale: an update and review. Personal. Individ. Differ. **47**(5), 385–395 (2009)
41. Terracciano, A., Lóckenhoff, C.E., Crum, R.M., Bienvenu, O.J., Costa, P.T.: Five factor model personality profiles of drug users. BMC Psych. **8**(1), 22 (2008)
42. Valeroa, S., Daigre, C., Rodríguz-Cintas, L., Barral C., Gomà-i-Freixanet, M., Ferrer, M., Casasa, M., Roncero, C.R.: Neuroticism and impulsivity: Their hierarchical organization in the personality characterization of drug-dependent patients from a decision tree learning perspective. Compr. Psychiatry **55**(5), 1227–1233 (2014)
43. Ventura, C.A., de Souza, J., Hayashida, M., Ferreira, P.S.: Risk factors for involvement with illegal drugs: opinion of family members or significant others. J. Subst. Use **20**(2), 136–142 (2014)
44. Yasnitskiy, L., Gratsilev, V., Kulyashova, J., Cherepanov, F.: Possibilities of artificial intellect in detection of predisposition to drug addiction. Perm University Herald Series “Philosophy Psychology Sociology” **1**(21), 61–73 (2015)
45. Zuckerman, M.: Behavioral Expressions and Biosocial Bases of Sensation Seeking. Cambridge University Press, New York (1994)